Log 708 - Chapter 6 Solutions

Halvard Arntzen

6.1

a. The formula takes a workers age (X_3), subtracts years of education (X_1) and preschool years (6). If a worker has been working continuously since end of education, this should give exactly the years of work experience X_1 that we want. For some workers, there may be a number of years after ending education that was not spent working. For such workers the formula over-estimates the true working experience. Given that the workers in the sample may have been in "working age" since about 1950 we should expect that a substantially larger proportion of the women have been at home raising kids etc. For these women, the true work experience can be greatly overrated by the formula.

b. The formula explicitly shows that one variable is a linear combination of the others. This is a direct violation of assumption E, and means that parameters can not be estimated for such model. We can read the data, and try a regression to see what we get from R in this case.

```
wagedata <- read.csv("M:/Undervisning/Undervisningh21/Data/Wages.csv")
head(wagedata)
```

```
##   obs  wage female nonwhite union education exper age wind femalenonw
## 1   1 11.55      1        0     0        12    20  38    1          0
## 2   2  5.00      0        0     0         9     9  24    0          0
## 3   3 12.00      0        0     0        16    15  37    1          0
## 4   4  7.00      0        1     1        14    38  58    0          0
## 5   5 21.15      1        1     0        16    19  41    1          1
## 6   6  6.92      1        0     0        12     4  22    1          0
```

```
failreg <- lm(wage ~ education + exper + age, data = wagedata)
coef(failreg)
```

```
## (Intercept)   education       exper         age
##  -9.5860816   1.4145209   0.1787117          NA
```

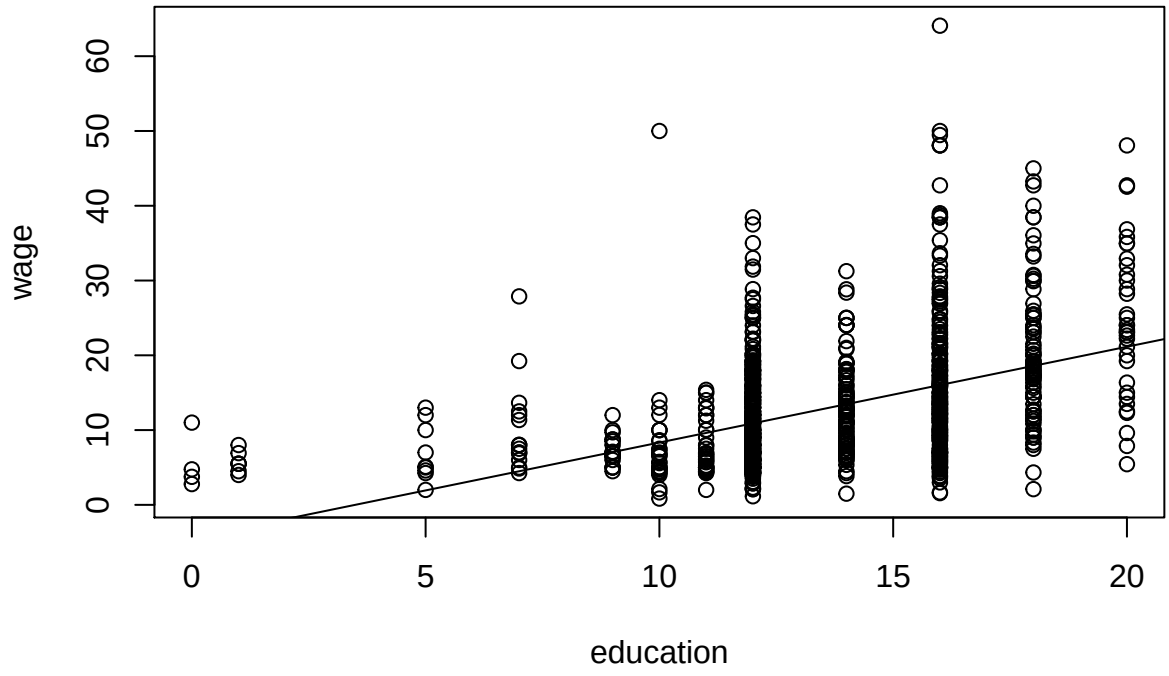
So, we see that R solves the problem by excluding the `age` variable, and then estimates the equation based on the two other variables. The coefficient for `age` is listed as `NA` (Not Available).

c. For model A, we're supposed to use only education.

```
#run regression
regA <- lm(wage ~ education, data = wagedata)
```

It's always a good idea to look at a scatterplot in this connection.

```
with(wagedata, plot(education, wage))
abline(regA)
```



There is some evidence of a positive correlation, although not very high.

Now, we can either look at the whole `summary` or "cherry-pick" what we want. So either

```
summary(regA)
```

```
##
## Call:
## lm(formula = wage ~ education, data = wagedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.486  -4.749  -1.399   3.101  48.057
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.4745     0.9356  -4.783 1.93e-06 ***
## education     1.2811     0.0696  18.408  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.028 on 1287 degrees of freedom
## Multiple R-squared:  0.2084, Adjusted R-squared:  0.2078
## F-statistic: 338.8 on 1 and 1287 DF,  p-value: < 2.2e-16
```

Where we find estimated equation $y = -4.47 + 1.28 \cdot x_1$ and $R^2 = 0.21$, or we can go

```
coef(regA)
```

```
## (Intercept)   education
##   -4.474476    1.281113
```

```
#compute, but "hide" summary in object s. Note that s will be a "list" that we can ins
#in the Rstudio "environment" window.
s <- summary(regA)

#get R square from s
s$r.square
```

```
## [1] 0.2084087
```

A 95% confidence interval is produced this way:

```
confint(regA)
```

```
##                 2.5 %    97.5 %
## (Intercept) -6.309881 -2.639071
## education    1.144577  1.417649
```

So, the interval is [1.14, 1.41].

d. So, let's run the regression first, either of the following codes will work:

```
#using full formula for model B
regB <- lm(wage ~ education + exper, data = wagedata)

#or using the update mechanism
regB <- update(regA, . ~ . + exper)
```

We can certainly answer the questions based on a `summary(regB)` code. That gives as follows

```
summary(regB)
```

```
##
## Call:
```

```
## lm(formula = wage ~ education + exper, data = wagedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.698  -3.928  -1.064   2.745  48.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.58608    1.00977  -9.493   <2e-16 ***
## education    1.41452    0.06770  20.894   <2e-16 ***
## exper        0.17871    0.01633  10.941   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.725 on 1286 degrees of freedom
## Multiple R-squared:  0.2758, Adjusted R-squared:  0.2747
## F-statistic: 244.9 on 2 and 1286 DF,  p-value: < 2.2e-16
```

However, since we are also going to compare model A and B, this is a good place to use the `stargazer` package:

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.0.3
```

```
stargazer(regA, regB, type = "text",
          ci = TRUE,
          keep.stat = c("n", "rsq"))
```

```
##
## ===============================================
##                       Dependent variable:
##              ---------------------------------
##                              wage
##                     (1)               (2)
## -----------------------------------------------
## education         1.281***          1.415***
##                (1.145, 1.418)    (1.282, 1.547)
##
## exper                               0.179***
##                                   (0.147, 0.211)
##
## Constant         -4.474***         -9.586***
##                (-6.308, -2.641) (-11.565, -7.607)
##
## -----------------------------------------------
## Observations      1,289             1,289
```

```
## R2                  0.208               0.276
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

This shows: For model B, both variables are highly significant (the P-values for the standard test is below 0.01 as indicated by "***"), both variables affect the wages positively, which is as expected. The R^2 increases to about 0.28 from 0.21, another sign that `exper` really adds explanatory power to the model A. The coefficient for `education` is not greatly affected by the inclusion of `exper`, as we can see by noting that the confidence intervals from model A, B are overlapping. A reason for this is likely that the two independent variables are not correlated in any substantial way. (In fact the correlation coeff is -0.18). (Recall from theory that if strong omitted variable bias in model A is caused by omitting `exper`, the two variables in model B would have to be more correlated.)

e. So, the coefficient estimates are *marginal effects* i.e, an extra year of education is estimated to cause on average 1.45$ increase in wage, while an extra year of experience gives on average 0.18$

f. The predictions can be computed by inserting the data for Maria and Jane into the estimated equation for model B. Furthermore, an approximate 95% error margin can be found as $2S_e$. We find S_e from R by either looking at the complete summary, or by doing:

```
s <- summary(regB)
s$sigma
```

```
## [1] 6.724914
```

So, the error margin can be taken at about 13.44.

We can be lazy and ask R to calculate the predictions with prediction interval.

```
newdata <- data.frame(education = c(12, 12),
                      exper = c(10, 15),
                      row.names = c("Maria", "Jane"))

pred_wages <- predict(regB, newdata, interval = "prediction")
cbind(newdata, round(pred_wages, 2))
```

```
##       education exper   fit   lwr   upr
## Maria        12    10  9.18 -4.03 22.38
## Jane         12    15 10.07 -3.13 23.27
```

The result shows that the model is not great for prediction, because of the wide intervals. Still the model *can* be good for estimating marginal effects!

g. Since model B does not include gender as a variable, it will not predict a different wage for Jim.

6.2

Ok, we read the file from our favorite place.

```
usedcars <- read.csv("M:/Undervisning/Undervisningh21/Data/used_cars.csv")
head(usedcars)
```

```
##       price age agecat mileage statwag newprice region eu_time corros report
## 1  69.11825  17      3     226       0      310      0       3      2      0
## 2  56.30901  16      3     186       0      270      0       5      2      0
## 3 141.90664   7      1      43       1      230      0      22      1      0
## 4 262.93979   2      1      45       0      320      0       0      0      1
## 5   9.00000  17      3     183       0      200      0      10      2      0
## 6 110.20784  10      2     117       0      240      1      15      2      1
```

We could get rid of the disturbing extra decimals by using `round(price, 2)` but let's not bother. We run the mentioned regressions as follows.

```
regA <- lm(price ~ mileage + newprice, data = usedcars)

regB <- lm(price ~ age + mileage + newprice, data = usedcars)

#or: regB <- update(regA, . ~. + age)

coef(regA)
```

```
## (Intercept)     mileage    newprice
## -21.0148029  -0.7820433   0.9284720
```

```
coef(regB)
```

```
## (Intercept)         age     mileage    newprice
##   5.1487847 -10.2882781  -0.1285636   0.8956327
```

So we get respectively

$$price = -21 - 0.78 \cdot mileage + 0.92 \cdot newprice$$

$$price = 5.15 - 10.29 \cdot age - 0.13 \cdot mileage + 0.89 \cdot newprice$$

for model A, B.

b. The estimate jumps from -0.78 to -0.13 when we also include age.

c. Since `age` and `mileage` is very correlated, when we only include mileage in model A this effect also covers a lot of the "age effect". When we actively include age in the model, we can isolate the effect of mileage so that the effect on cars of same age can be estimated.

d. None of them are in fact wrong, but the parameter in question is not actually the same effect as described above. That's why we get different intervals. However, we can say that -0.78 is a wrong estimate for the *marginal effect* of `mileage`.

6.3

Suppose prices on average drop by p percent per year of age, and by q per added 1000km of mileage. Then a model for the price Y, given X_1, X_2, X_3 as in exercise **??** is as follows.

$$Y = X_1 \cdot \left(1 - \frac{p}{100}\right)^{X_2} \cdot \left(1 - \frac{q}{100}\right)^{X_3} + \mathcal{E} \ .$$

This is a highly nonlinear model. We will learn in subsequent chapters how to estimate parameters for this and other types of non-linear models. To clean up a little bit: It will often be more convenient to model errors as multiplicative rather than additive for such models. Also, with

$$r = \left(1 - \frac{p}{100}\right), \quad s = \left(1 - \frac{q}{100}\right) \ ,$$

we can reformulate the model more compactly as

$$Y = X_1 \cdot r^{X_2} \cdot s^{X_3} \mathcal{E} \ .$$

The interesting parameters to estimate here would be r and s. For example, an r at 0.88 would imply an average value loss of 12 % per year of age for used cars.